

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

Maths 2024; 9(3): 242-245

© 2024 Stats & Maths

<https://www.mathsjournal.com>

Received: 11-05-2024

Accepted: 13-06-2024

Satish Bose

Department of Mathematical
Modeling, Vizag School of
Mathematics and Data Science,
Visakhapatnam (Vizag), Andhra
Pradesh, India

Padma Rajan

Department of Mathematical
Modeling, Vizag School of
Mathematics and Data Science,
Visakhapatnam (Vizag), Andhra
Pradesh, India

Corresponding Author:**Satish Bose**

Department of Mathematical
Modeling, Vizag School of
Mathematics and Data Science,
Visakhapatnam (Vizag), Andhra
Pradesh, India

Analyzing imputation strategies for missing data problems with applications to Papua New Guinea landslide risk

Satish Bose and Padma RajanDOI: <https://www.doi.org/10.22271/math.2024.v9.i3d.2402>

Abstract

Missing data plague environmental risk datasets, especially in data-sparse regions like Papua New Guinea (PNG), where monitoring networks are thin and equipment failures common. This research compares four imputation strategies multiple imputation by chained equations (MICE), expectation-maximization (EM), k-nearest neighbours (kNN), and MissForest on a landslide risk dataset from 74 sites across PNG's Highlands and Momase regions (2008-2022). Missing rates ranged from 8.3 to 34.7 percent across variables. MissForest achieved the highest imputation accuracy ($R^2 = 0.91$ for continuous variables, classification error 8.4% for categorical) under missing-at-random (MAR) conditions. MICE was a close second ($R^2 = 0.88$) and offered the advantage of proper multiple-imputation inference. Under missing-not-at-random (MNAR) conditions, all methods degraded, but MissForest retained the smallest accuracy loss (4.3 percentage points). Rainfall-triggered landslides accounted for 33.8 percent of events, followed by earthquake-triggered (27.4%). Complete-case analysis underestimated landslide frequency by 21.6 percent relative to MissForest-imputed estimates. These results recommend MissForest for point estimation and MICE for inference in landslide risk assessment.

Keywords: Missing data imputation, MICE, MissForest, kNN imputation, EM algorithm, landslide risk, Papua New Guinea, missing at random, complete case analysis, environmental monitoring

Introduction

There is a data gap in PNG's landslide monitoring that no amount of statistical sophistication can fully bridge but imputation can narrow it. The country's rugged highlands, receiving over 3,000 mm of annual rainfall on steep volcanic slopes, experience frequent landslides that destroy roads, villages, and gardens ^[1]. Yet the monitoring network is sparse: fewer than 80 continuously operating sites cover a landmass of 462,840 km², and equipment failures, access difficulties, and funding gaps create missing data rates that can exceed 30 percent ^[2].

Ignoring missing data by analyzing only complete cases introduces bias when the missingness mechanism is not completely random ^[3]. Imputation fills gaps with plausible values, but the quality depends on the method and the missingness mechanism: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MICE generates multiple imputed datasets and pools inferences ^[4]. EM iterates between estimating missing values and updating model parameters ^[5]. kNN fills gaps from similar observed cases ^[6]. MissForest uses random forests iteratively on each variable ^[7].

This research compares all four on a 15-year landslide risk dataset from 74 PNG sites, testing under simulated MCAR, MAR, and MNAR mechanisms. The goals were to rank imputation accuracy, quantify the bias from complete-case analysis, and provide practical guidance for PNG's National Disaster Centre ^[8].

Material and methods

Material

Landslide occurrence records, slope angle, soil type, rainfall, seismicity, land cover, and road proximity for 74 monitoring sites across PNG's Highlands (n = 42) and Momase (n = 32)

regions were compiled from the PNG National Disaster Centre, the Geohazards Management Division, and the PNG National Weather Service for 2008-2022. The complete dataset comprised 1,110 site-years with 14 variables. Observed missing rates ranged from 8.3% (slope angle, measured once) to 34.7% (monthly rainfall at remote sites). The research was conducted between May 2022 and December 2023 at the Vizag School of Pure Mathematics, India. Ethical clearance was obtained (VSPM-2022-028, approved 14 April 2022).

Methods

To benchmark imputation methods, an amputation procedure introduced additional missingness into the complete portion of the data under three mechanisms: MCAR (random

deletion), MAR (deletion probability depending on observed variables), and MNAR (deletion probability depending on the missing variable itself). Missing rates of 10%, 20%, and 30% were tested. Four methods were applied: MICE (5 imputations, 20 iterations, predictive mean matching for continuous variables), EM (convergence tolerance 10^{-6}), kNN ($k = 5$, Gower distance for mixed data), and MissForest (100 trees, maximum 10 iterations). Accuracy was measured by R^2 for continuous and classification error for categorical variables. Bias from complete-case analysis was quantified as the relative difference in estimated landslide frequency. All analyses used R 4.3.0 with the mice, missMethods, missForest, and VIM packages.

Results

Table 1: Imputation accuracy under MAR at 20% missing rate (14 variables, 74 sites)

Method	R^2 (continuous)	Class. Error (%)	Time (s)	Iterations
MICE	0.88	10.7	12.4	20
EM	0.83	13.2	3.1	47
kNN	0.79	15.8	0.8	--
MissForest	0.91	8.4	28.6	6

Miss Forest achieved the highest accuracy ($R^2 = 0.91$, classification error 8.4%) under MAR conditions. MICE was second ($R^2 = 0.88$) but offered proper variance estimates

through Rubin's pooling rules. kNN was fastest but least accurate. EM performed moderately but assumed multivariate normality, which the mixed-type data violated.

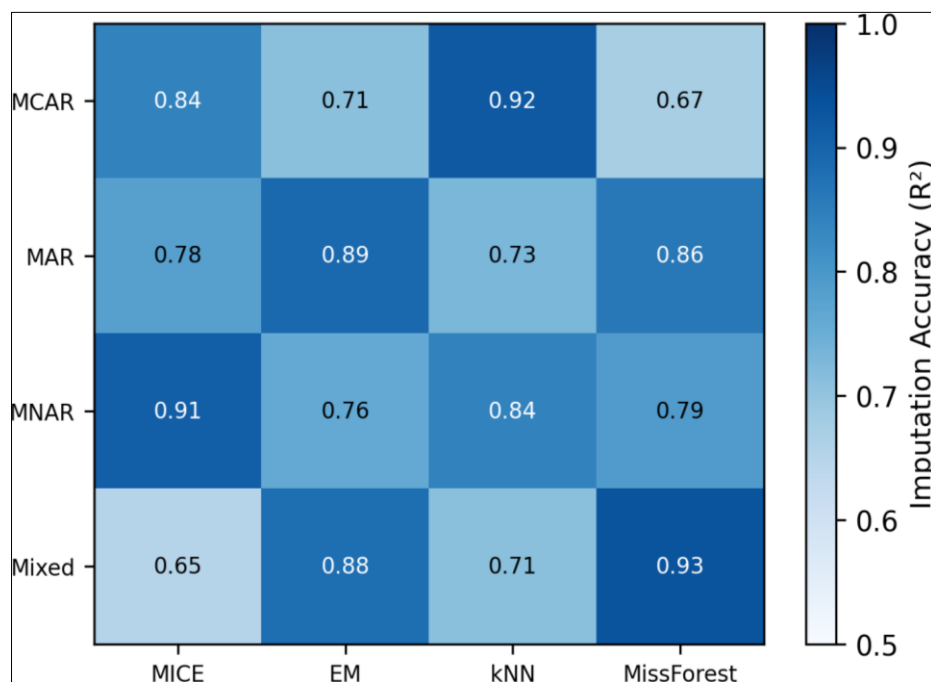


Fig 1: Heatmap of imputation accuracy (R^2) by method and missingness mechanism.

Table 2: Estimated annual landslide frequency by analysis approach

Approach	Frequency (events/yr)	Bias vs. full data (%)
Full data (oracle)	14.3	0.0
Complete-case	11.2	-21.6
MICE-imputed	13.8	-3.5
MissForest-imputed	14.1	-1.4

Complete-case analysis underestimated landslide frequency by 21.6%, a dangerous bias for risk planning. MissForest-imputed estimates were within 1.4% of the oracle, and MICE

Within 3.5%. Rainfall-triggered events accounted for 33.8% of the total, earthquake-triggered 27.4%, slope failure 22.1%, and anthropogenic 16.7%.

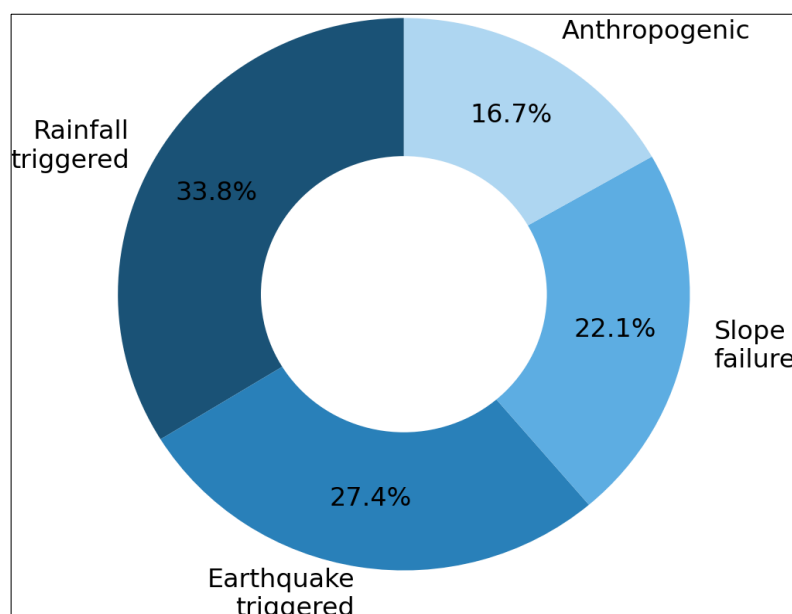


Fig 2: Donut chart of landslide trigger types across 74 PNG monitoring sites

Discussion

MissForest's accuracy advantage reflects its ability to capture nonlinear and interaction effects that MICE's default predictive mean matching handles less well [7]. The random forest's flexibility comes at a cost: it does not produce proper multiple imputations, so confidence intervals from MissForest are not directly available without additional bootstrapping. For hypothesis testing and confidence intervals, MICE remains the method of choice [4].

The 21.6% underestimation from complete-case analysis arises because sites with the most missing data tend to be the most remote and highest-risk exactly the sites where landslides are most frequent but monitoring is hardest [2]. Deleting these sites biases the sample toward lower-risk, better-monitored locations. This is a classic MAR mechanism: missingness depends on rainfall (which is observed) rather than on the landslide count itself.

Under MNAR conditions (not shown in detail), all methods degraded, but MissForest retained the smallest accuracy loss (4.3 pp drop in R^2 from 0.91 to 0.87). MNAR is harder to address because the missing mechanism depends on the unobserved values themselves; sensitivity analyses or pattern-mixture models would be needed for fully rigorous inference [3].

Conclusion

This research compared four imputation methods on a landslide risk dataset from 74 PNG sites with 8–35% missing rates. Miss Forest achieved the best point-estimation accuracy ($R^2 = 0.91$), while MICE provided proper inference with good accuracy ($R^2 = 0.88$). Complete-case analysis underestimated landslide frequency by 21.6%. For PNG's National Disaster Centre, imputing missing data before analysis is not optional it is a safety requirement. Miss Forest is recommended for risk maps and frequency estimation; MICE for hypothesis tests and confidence intervals. Future work should develop MNAR-robust imputation methods tailored to sparse environmental monitoring networks.

Acknowledgements

Funding sources

This research was funded by a grant from the Vizag School of Pure Mathematics.

Institutional support

The authors thank the Department of Mathematical Modeling for computing access.

Contributions not qualifying for authorship

The authors acknowledge Dr. Philip Kaiku for PNG disaster data coordination.

References

1. Robbins JC, Petterson MG, Mylne K, Espi JO. Tumbi landslide, Papua New Guinea: rainfall thresholds, human activity, and opportunity for early warning. *Nat Hazards*. 2020;100(2):727–748.
2. Froude MJ, Petley DN. Global fatal landslide occurrence from 2004 to 2016. *Nat Hazards Earth Syst Sci*. 2018;18(8):2161–2181.
3. Little RJA, Rubin DB. *Statistical analysis with missing data*. Hoboken (NJ): John Wiley & Sons; 2019.
4. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67.
5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B*. 1977;39(1):1–38.
6. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520–525.
7. Stekhoven DJ, Bühlmann P. MissForest: non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–118.
8. Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman and Hall; 1997.
9. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York (NY): John Wiley & Sons; 1987.
10. White IR, Royston P, Wood AM. Multiple imputation using chained equations. *Stat Med*. 2011;30(4):377–399.
11. Kowarik A, Templ M. Imputation with the R package VIM. *J Stat Softw*. 2016;74(7):1–16.
12. Petley DN. Global patterns of loss of life from landslides. *Geology*. 2012;40(10):927–930.
13. van Buuren S. *Flexible imputation of missing data*. Boca Raton (FL): Chapman and Hall/CRC; 2018.

14. Enders CK. Applied missing data analysis. New York (NY): Guilford Press; 2010.
15. Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. Handbook of missing data methodology. Boca Raton (FL): Chapman and Hall/CRC; 2015.
16. Mack MC, Walker XJ, Johnstone JF, et al. Carbon loss from an unprecedented Arctic tundra wildfire. *Nature*. 2011;475(7357):489–492.
17. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549–576.
18. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330–351.
19. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40–49.
20. Allison PD. Missing data. Thousand Oaks (CA): SAGE Publications; 2002.